Detection and Segmentation to locate and identify objects of interest and eye movement patterns in VR art exhibitions

BATUHAN USTA, University of Twente, The Netherlands

Virtual Reality (VR) integration has gained momentum in today's technological era, especially in museums, aiming to enhance visitor engagement. Identifying the objects that interest visitors and based on these interests providing information to engage with visitors in VR environments, particularly in art museums, poses challenges. This paper proposes leveraging deep learning models, particularly object detection models incorporated with segmentation models, to accurately detect and locate objects in paintings within VR environments to identify objects of interest. This research contributes to advancing VR applications in art museums and potentially other sectors requiring accurate object detection in immersive environments. Additionally, the study explores methods for comparing pixel locations of segmented sections with the eye gaze information gathered from the VR environment to obtain the patterns of the visitor's eye movements.

Additional Key Words and Phrases: Semantic Segmentation, Panoptic Segmentation, Virtual Reality, Eye Gaze, Segment Anything Model, Pseudo-Paintings

1 INTRODUCTION

In today's rapidly evolving technological landscape, Virtual Reality (VR) is increasingly being integrated into various sectors, including museums [1, 15, 17, 28]. This adoption is partly driven by a significant decline in art museum visitors, highlighting the need for technologies like VR to boost engagement [25]. VR offers an immersive, educational, and entertaining experience, allowing visitors to explore museum exhibits interactively. Additionally, virtual agents can provide personalized information, enhancing visitor engagement and attention [23]. Museums can use VR to create captivating systems for presenting information, such as interactive mechanisms that display detailed information about artworks [2, 7].

Furthermore, VR museums have the advantage of mobility which overcomes the barrier of distance since the system does not require a destination and simply can be displayed from anywhere as long as they have the hardware which supports it [15]. This advantage allows multiple artworks to be showcased at the same time since the system does not have physical storage restrictions. However, the data storage need arises from this advantage. One feasible system which can serve this need is to use a knowledge graph [12]. The knowledge graph has multiple use cases for the VR environment. They can be used to store the displayed artworks and their information for example, if the artwork is a painting, then the painter and the description of the painting can be stored. Besides, museums can also utilize the knowledge graph to store the location of objects within the artworks. This stored information is very useful since it can be used to identify objects of interest which is engrossing because it can help to attract users' attention [2, 7, 27]. However, it is challenging to identify what objects interest visitors [3].

One promising way of addressing that challenge would be to gain insight into the eye gaze data of the visitors. This data offers insights into what captures the visitors' interest by locating the visualization point of the user [19]. This information can be stored in the knowledge graph along with the location of the objects. Moreover, the location of the object and the eye gaze data can be compared to determine which object corresponds to the coordinates the visitor is looking at. However, to compare the two data, the knowledge graph must encompass every object depicted in the artwork, along with their respective locations and relevant information.

Manual annotation of objects in artworks is time-consuming and inefficient, especially with frequent updates. To address this, we utilize Deep Learning (DL) models for object detection [6, 8, 20, 33]. Extensive research identifies Faster R-CNN and YOLO as the top DL models for this purpose [16, 22]. These models detect objects and define their boundaries, outputting a list with a class label and bounding box coordinates. While useful for general object detection, plotting bounding boxes is unnecessary in VR art museums, where the goal is to identify objects of interest. Additionally, bounding boxes can cause overlap, as they do not precisely represent object edges.

Segmentation models address challenges in identifying pixel locations of objects, differing from object detection models which focus on bounding boxes. Popular segmentation models include Mask R-CNN, Panoptic Segmentation Model, YOLOv8 Instance Segmentation, F-SAM, and UPSNet [10, 21, 26, 29, 32]. However, their accuracy is often limited due to small training datasets [4]. The Segment Anything Model (SAM) offers higher accuracy, trained on 11 million images with 1 billion masks, though it lacks class labels [14]. A solution is auto-annotation, which uses object detection to provide bounding boxes and class labels, narrowing the search area for the segmentation model to label the masks produced by SAM accurately.

The VR art museums can use the pre-trained object detection and segmentation models to auto-annotate paintings. However, there is an issue with using the pre-trained object detection models. The pre-trained object detection models have limited and fixed class labels and some of these labels do not provide any relevant information in paintings. Additionally, due to the limited class labels, some important objects are overlooked and not detected in the process. One solution could be to use custom datasets that include the missing essential class labels. These datasets can be created by manually collecting and annotating images or by utilizing annotated image collections from open sources [31]. We propose to use both options to obtain the most suitable dataset to detect objects in paintings.

Another challenging factor is the pixel quality and the brightness of the paintings [9, 18, 30]. State-of-the-art models are trained and tested on bright and quality images and according to the research,

TScIT 41, July 5, 2024, Enschede, The Netherlands

the existing models perform poorly on dark and low-pixel-quality images [18, 24]. One way of solving such an issue is collecting a dataset just containing paintings and annotating them manually [4]. However, this would not be possible due to the limited open-source paintings. Therefore, according to the research, converting images within the existing dataset to dark and low quality is suggested [9, 30]. We propose to use a pre-trained Neural Style Transfer Network (NSTN) to convert the collected images into more suitable versions to increase the accuracy when performing on the paintings [9].

We will be using the following research questions (RQ) as the basis of our research:

- **RQ1:** What are the most effective deep learning models and methods for accurately segmenting and annotating pixel locations of objects in classical paintings?
  - **RQ1.1:** How does the expansion of class labels influence the accuracy and comprehensiveness of deep learning models in segmenting and annotating objects in classical paintings, and what significant objects might be missed due to a limited set of labels?
  - **RQ1.2:** Does style transfer offer advantages or disadvantages for object detection in classical paintings?
- **RQ2:** How can segmented and annotated objects from classical paintings be utilized in cooperation with the eye gaze information to provide valuable data?

In this research, we analysed classical paintings and their descriptions to identify important class labels and determine the objects. We then collected a dataset based on these class labels. Additionally, we used a style transfer method to create a visually identical but stylistically different dataset to compare the advantages and disadvantages of using styled datasets for object detection on paintings. We trained two detection models on the custom datasets and evaluated them to assess the effects of styling. We integrated the SAM model with the object detection model to auto-annotate paintings, obtaining accurate segmentation masks. These masks were used to analyse eye gaze information and determine the eye movement patterns of visitors. Finally, these patterns were analysed to verify if the defined class labels were as important as described.

Our results showed that the model trained on the non-styled dataset outperformed the styled dataset model. We integrated the SAM model with the object detection model to auto-annotate paintings, obtaining accurate segmentation masks. The auto-annotation system functioned as expected, demonstrating high accuracy. These masks were used to analyse eye gaze information and determine the eye movement patterns of visitors. Finally, these patterns were analysed to verify if the defined class labels were as important as described, and the results confirmed their significance.

2 RELATED WORK

In this section, we will review recent advancements and key works in the following areas: the Segment Anything model, object detection models including Faster R-CNN and YOLOv8, the limitations of detection models on paintings, style-transfer methods, knowledge graphs and VR art museums.

2.1 VR Art Museums

Virtual Reality (VR) in existing systems has surged in popularity across various sectors [1, 17, 28]. Museums are one of the most popular sectors for the usage of VR [15]. The rapid adaptation of VR technologies into museums was crucial due to the significant decrease in visitor numbers for art museums [25]. This decrease emphasized the importance of utilizing technologies such as VR to enhance visitor engagement [25]. VR provides an educational, entertaining, visionary, and engaging environment in which visitors can be fully immersed [15]. Furthermore, the VR environment provides the opportunity to include a virtual agent for guidance to visitors. The virtual agents are inspired by the tour guides who play an essential role in museums, they offer extra information tailored to visitors' interests to increase engagement and maintain attention [2, 7]. Recent studies indicate that providing information about art or objects of interest increases visitors' engagement and attention [23]. Museums can integrate engaging and captivating systems to present information to their visitors [7]. In this open era, there are multiple ways to showcase artworks aesthetically while engaging audiences [15]. For instance, one approach could involve incorporating an interactive mechanism like a button to control the information visualization system.

2.2 Object Detection

Object detection is a computer vision task to identify and locate various objects within an image. The location of the object is usually presented with a bounding box which is identified by the pixel locations of its four edges. The bounding box itself does not provide the full information and therefore, the output is produced with its label attached to it. The label provided with the bounding box is called the class label. The class label provides the identity of the object by categorizing them into various classes (i.e., man, woman, animal) [8].

According to research, there are multiple object detection models used in the industry [16]. The most popular models include Faster R-CNN and YOLO models [22]. It is not possible to determine which of the two models is more efficient, as their usefulness depends on the specific conditions and applications for which they are used. Faster R-CNN (Region-based Convolutional Neural Networks) is a two-stage object detection model. It consists of a region proposal network (RPN) and a detection network that classifies the proposed regions and refines their boundaries. On the other hand, YOLO is a single-staged object detection model that simultaneously predicts bounding boxes and class probabilities for multiple objects in an image. Due to the additional stage in the Faster R-CNN model, it achieves higher accuracy compared to the YOLO model [5]. However, the YOLO model compensates for its lower accuracy with greater speed. For custom training, similar accuracy can be achieved by training YOLO on more data within the same time frame [22].

2.3 Image Segmentation

Image segmentation is a method which provides the ability to identify and differentiate various objects in images [13, 21]. Traditional segmentation methods usually require extensive manual input to work accurately. However, with the advances in AI and deep learning models like fully convolutional neural networks (FCNN), the segmentation methods became more autonomous. These methods still face challenges in terms of accuracy and efficiency [9]. Training segmentation models is a time and effort-consuming process and the existing models do not perform accurately on unidentified objects. Some of the most recent segmentation models are the following:

- Mask Region-Based Convolutional Neural Network (R-CNN) [10]
- Panoptic Segmentation Model [13]
- You Only Look Once Version 8 (YOLOv8) Instance Segmentation [32]
- Fast Segment Anything Model (SAM) [21]
- A Unified Panoptic Segmentation Network (UPSNet) [29]

These models are being used in the market for performing segmentation tasks. However, due to the limitations in annotated training images for such models, they have low accuracy unless provided with manually annotated images, which is time-consuming. There is another segmentation model in the market that overcomes the accuracy problem: the Segment Anything Model (SAM). The SAM model is an advanced computer vision model designed to segment various objects in an image. It does not provide the distinct categories that the segmented objects belong to but rather classifies everything in a single category. By eliminating the traditional classification step, SAM directly segments objects in images, leading to higher accuracy and versatility across various types of objects and scenes [14]. Accuracy is crucial for detecting and locating objects in this research. Therefore, the Segment Anything Model (SAM) has been chosen for use in this paper.

2.4 Style-Transfer

Object detection and segmentation tasks are more difficult to perform on paintings due to their diversity in terms of both colour and shapes of objects, and backgrounds. Such diversity causes models to perform poorly [4]. The sole cause of this problem is related to the fact that the existing models are trained on photographic datasets, and they do not include enough painting images to balance the bias. Style transfer methods are used to overcome this challenge. Style-transfer method is a technique in the field of computer vision and digital image processing where the artistic style and the content of an image blend into the target image. This process allows for the inclusion of diverse colours, textures, and styles in a dataset [4, 24, 30]. Style transfer methods influence various characteristics in an image and different methods focus on different characteristics while transferring the style. In the case of classical paintings, colours and pixel quality are more important than other characteristics. Therefore, we use our own style transfer model specifically focusing on the colour and the pixel quality of the images.

2.5 Knowledge Graphs

Knowledge graphs are a structured form of knowledge representation that captures information in a graph format, comprising nodes, edges, and properties. They provide a flexible and intuitive way to organize and query data, enabling powerful knowledge discovery and reasoning capabilities [12]. This method provides a more suitable environment for storing segmented and annotated information about objects due to the flexibility it provides while storing complicated information. Conventional storage uses table formatting which requires to categorize and structure the data to be stored. It is not possible to categorize and structure these pieces of information into a single column due to their complexity and variety. Thus, a knowledge graph serves as a great tool to be used in the case of storing location information of identified objects and their pixel coordinates since it does not require a structure or a category but rather utilizes the labels and connections of the data.

3 METHODOLOGY

This section will explain the steps we take in order to find an answer to the research questions introduced. The steps taken during the research are structured and identified in Figure 1.

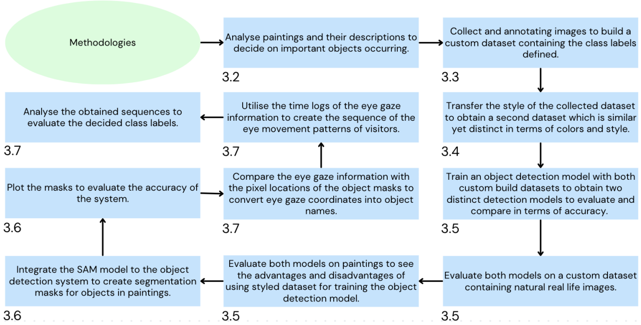

Methodologies
Analyse paintings and their descriptions to decide on important objects occurring.
3.2
Collect and annotating images to build a custom dataset containing the class labels defined.
3.3
Analyse the obtained sequences to evaluate the decided class labels.
3.7
Utilise the time logs of the eye gaze information to create the sequence of the eye movement patterns of visitors.
3.7
Transfer the style of the collected dataset to obtain a second dataset which is similar yet distinct in terms of colors and style.
3.4
Plot the masks to evaluate the accuracy of the system.
3.6
Compare the eye gaze information with the pixel locations of the object masks to convert eye gaze coordinates into object names.
3.7
Train an object detection model with both custom build datasets to obtain two distinct detection models to evaluate and compare in terms of accuracy.
3.5
Integrate the SAM model to the object detection system to create segmentation masks for objects in paintings.
3.6
Evaluate both models on paintings to see the advantages and disadvantages of using styled dataset for training the object detection model.
3.5
Evaluate both models on a custom dataset containing natural real life images.
3.5

Fig. 1. Diagram of the steps taken during the methodologies.

3.1 Object Detection Model

YOLOv8 is used for this research as it provides the most accurate results considering the limited time for training. The library provides pre-trained models as well as the option to custom-train them. Custom training is the same process as fine-tuning. Fine-tuning is an excellent option given the time constraints, as it builds upon an existing detection model, thereby skipping the initial model-building process. Additionally, fine-tuning provides the option to define custom class labels due to their importance in capturing meaningful objects in paintings. In this research, the YOLOv8 model is fine-tuned on two custom-created datasets.

3.2 Determining Important Class labels

To identify the meaningful class labels in classical paintings, manual detection has been done. The paintings and their descriptions are analyzed in terms of mostly occurring objects. The first step taken in this process was to determine the common words occurring in descriptions of the paintings. These common words do not provide all the important objects occurring in the paintings since the descriptions miss some of the objects due to their irrelevance to history or the exhibit that they are presented in, Secondly, the paintings are observed, and additional classes are included by looking at their occurrence counts. After these two steps, 11 meaningful class labels

Table 1. The number of collected images per class label in the normal and styled training dataset and normal validation dataset.

	Man	Woman	Animal	Building	Vase	Hat	Necklace	Clothing	Table	Sword	Belt
Train	4000	3800	4000	4000	187	882	138	3817	583	100	40
Val	292	204	74	128	21	42	19	357	173	9	2
Total	4292	4004	4074	4128	208	924	157	4174	756	109	42

are determined to be important and relevant to include in the detection system. These labels and their occurrence counts are presented in Table 2.

3.3 Creating Custom Dataset

OIDv4 was used to collect a dataset containing some of the 11 class labels identified from the analysis of paintings. These classes include man, woman, animal, building, table and clothing. The total number of images collected from the open-source dataset is 20,200 for training and 1,228 for validation. Since the open-source dataset does not cover all the required classes, the remaining 5 classes are gathered by manually collecting images. Additionally, Roboflow was utilized to manually label these images, which were then added to the collected dataset to obtain a custom-created dataset for training. The number of images collected for training and validating models per class label can be seen in Table 1. These images are natural real-life images and therefore they do not resemble paintings causing them to be a biased dataset. However, to decrease the bias of the dataset, augmentation is applied to the images. The augmentation process simply includes adding noise into the image as well as adjusting the brightness to obtain a more unbiased dataset. Three versions of images are created with the augmentation step. The first version is the original version of the image while the second version includes noises in the image. The noises are simply the black pixel spots randomly distributed throughout the image. Lastly, the third version includes noises and the adjusted brightness. The brightness is increased by 2.3%. An example of three versions of augmented images can be seen in Figure 2.

3.4 Styling Custom Dataset

The research compares the advantages and the disadvantages of styled images in the training process; therefore, the collected dataset for training was fed into a style-transfer model to obtain a similar but stylistically transformed dataset. This transferred dataset contains the same images gathered for the first dataset. The model trained on the styled dataset is expected to be more accurate for detecting objects in paintings since the styled dataset resembles in style and colour of the paintings. An example sample of normal and style-transferred images can be seen in Figure 3. By evaluating the performance of the models on both datasets, we aim to understand the impact of style transfer on object detection accuracy. This analysis will help determine whether incorporating styled images into the training process provides a significant benefit for detecting objects in classical paintings.

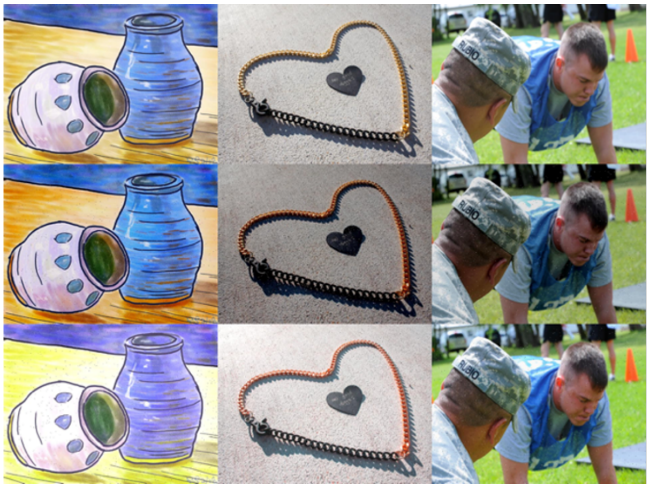


Fig. 2. Sample of the three versions of augmented images.

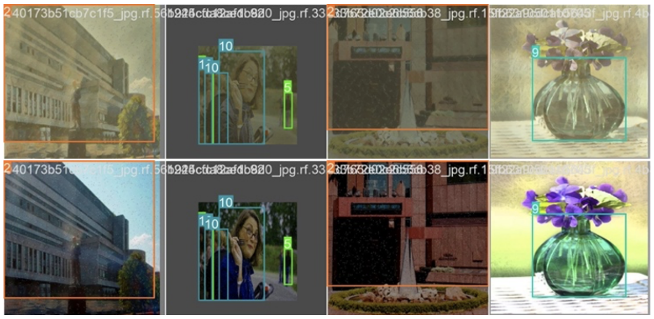


Fig. 3. Sample of the normal and styled images during the training process.

3.5 Evaluation of Models

Evaluation of the styled and non-styled models was conducted by utilizing the features provided by the Ultralytics library, which provides various metrics to compare and analyze the results of the models. For this research, we focused on the Normalized Confusion Matrix (NCM) and the F1-Confidence Curve. The NCM summarizes the comparison between the predicted and actual objects in the paintings, making it a valuable metric for this research as it shows how well the models performed on both the images and the paintings. Additionally, the F1-Confidence Curve represents the accuracy of the models by incorporating both precision and recall into a single graph. The F1-Confidence Curve is used to compare the consistency of models in terms of accuracy by analysing the different thresholds. The graph represents the thresholds as its x-axis and the F1 score as

Table 2. Object list with the number of occurrences in paintings.

Objects	**Man**	**Woman**	**Clothing**	**Hat**	**Building**	**Animal**	**Table**	**Vase**	**Necklace**	**Belt**	**Sword**
Counts	15	12	15	7	10	4	5	2	3	1	1

its y-axis. The threshold shows that if you choose a confidence level represented by the threshold, you simply decide how confident the detection has to be. If the threshold is 0.5, the model must be more confident than the threshold to detect the object, indicating that an object exists in that section of the image.

The two trained models were evaluated on both the collected validation dataset and the paintings for comparison. The results of the NCM and F1-Confidence Curve were collected to assess the performance.

3.6 Auto-Annotation

To obtain pixel-level detection, the Segment Anything Model (SAM) was incorporated into the system. The results of the object detection models were fed into the SAM model to generate masks for the labels. A mask is essentially a list of pixel locations representing the detected object. Moreover, the masks are plotted onto the images to manually check if the model's accuracy is sufficient for further analysis. Additionally, these masks enabled us to convert eye gaze information into object labels for further analysis of viewing patterns. These patterns represent the sequence of objects that visitors look at in a painting.

3.7 Eye Gaze Comparison

The collected eye gaze data is stored in CSV format and the data includes timestamps, painting name and the x-y coordinates of the gaze. In this research, these data are compared with the masks created from the auto-annotation process. The comparison process is divided into sections by examining a single painting for a single visitor session at a time. This process is repeated for all paintings and visitors to get a general overview. A single visitor session and a single painting mask text file are fed into the system at a time. The information is compared by first filtering the rows in the CSV file representing the visitor session by the painting name to increase speed. Secondly, the filtered rows are iterated, and the coordinates are compared with the entire text file, which contains the pixel values of the masks. If the pixels match one of the pixel coordinates of a mask, the timestamp and the object label are stored in a variable until the object label of the matched pixel coordinates is different from the previously matched one or if the pixel does not match any of the pixels in the masks, indicating the visitor is looking at the background. After going over all the rows, a sequence text is created for readability. The text involves the name of the object and the duration of the gaze on that object. Furthermore, the sequences are analysed to identify the important objects in paintings. This information is used to see which objects are important for visitors.

4 RESULTS

The results section is structured into four parts to address the research questions comprehensively. The first part gives an overview of the materials and the tools used while conducting this research. The second part presents the analysis of objects occurring in paintings, highlighting the importance of selecting appropriate classes for detection. Moving forward, the third part discusses the results from validating the styled and non-styled models on the validation set and the paintings to compare their accuracies. Lastly, the fourth part examines the eye movement patterns of visitors by presenting the segmentation results and the comparison results accordingly. These results are analysed to have a further understanding of which object class is more important.

4.1 Materials

The research tasks were conducted using a MacBook Pro equipped with a 2 GHz Quad-Core Intel Core i5 processor. The ANACONDA[1] environment, along with Jupyter notebooks, facilitated the utilization of various Python libraries, which were integral to the development and execution of the code. For dataset acquisition, OID v4 (Open Images Dataset v4)[2] was employed to gather datasets specifically tailored for custom model training. Roboflow[3] was used to manually label and integrate additional images collected for the dataset. Additionally, a style-transfer model was utilized to augment and diversify the dataset for comparative analysis. For training and evaluation purposes, the Ultralytics[4] library was chosen due to its comprehensive built-in functionalities, particularly for evaluation metrics. The 19 paintings used for testing and evaluation of detection models, along with the eye gaze data for analyzing eye movement patterns of visitors, were gathered from a separate study. This study involved 31 participants and aimed to gain insights into their experience of the VR museum [11].

4.2 Important Class Labels

The count of objects occurring in the paintings is shown in Table 2. The table lists the counts for each class separately, focusing only on the 11 classes with the highest counts. As evident from the values, the distinction between "man" and "woman" is the most significant, as these terms have the highest counts. This is also reflected in the descriptions, which are predominantly describing the portraits of people. The importance of additional class labels is analysed by using different pre-trained datasets and observing the number of detected objects. When two pre-trained models are tested on the paintings, results show that the pre-trained model including these 11 additional class labels detected 70% more objects than the pre-trained model excluding the additional 11 classes. Moreover, the model which does not have the additional 11 classes lacked the distinction between important labels like "man" and "woman".

[1] https://www.anaconda.com/

[2] https://github.com/EscVM/OIDv4_ToolKit

[3] https://roboflow.com/

[4] https://yolov8.com/

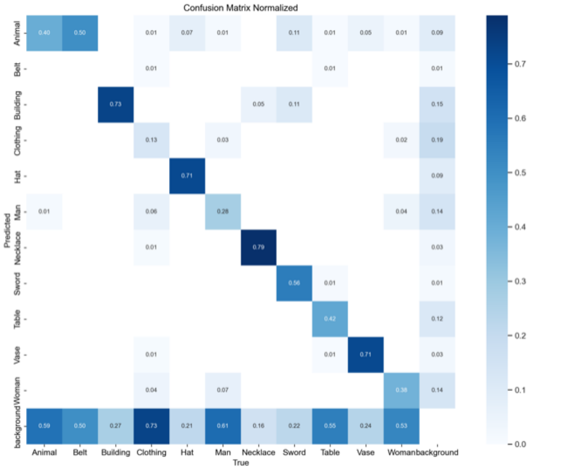


Fig. 4. Normalized Confusion Matrix results of the non-styled object detection model.

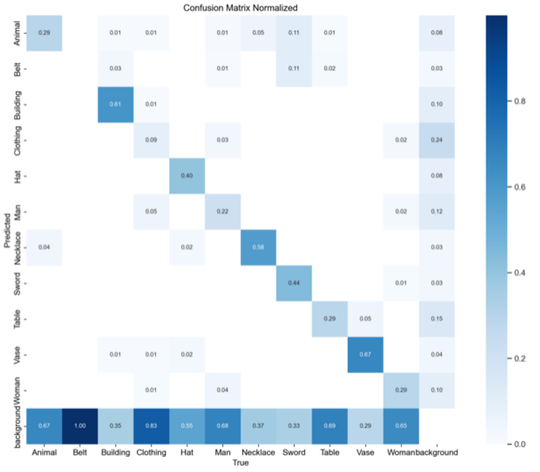


Fig. 6. Normalized Confusion Matrix results of the styled object detection model.

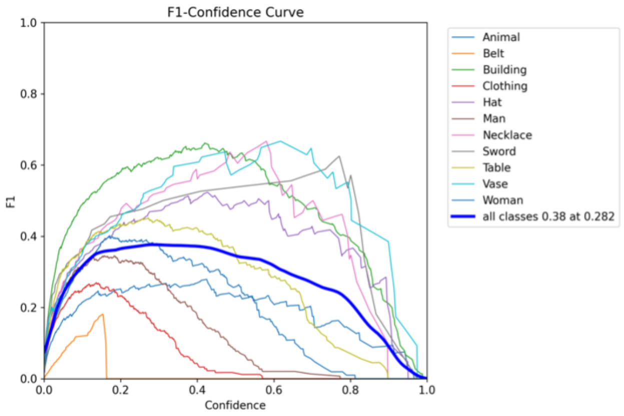


Fig. 5. F1- Confidence Curve results of the non-styled model.

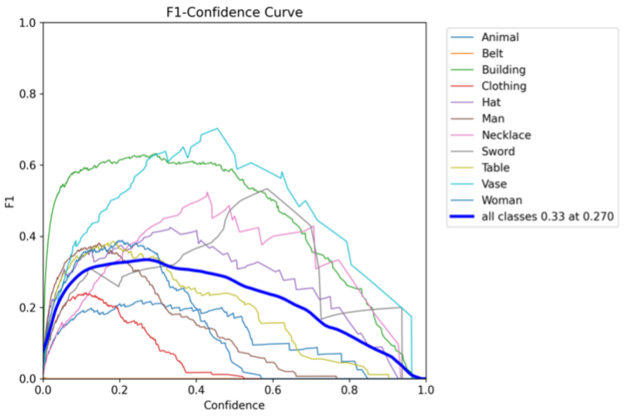


Fig. 7. F1- Confidence Curve results of the styled model.

4.3 Validation

The normalised confusion matrix gives insight into the accuracy of the predictions as it breaks down the results into categories such as true positives, true negatives, false positives, and false negatives, allowing for a deeper understanding of where the model succeeded and where it struggled. The analysis of these normalized confusion matrices for both the styled and non-styled models indicates that the styled model exhibits lower accuracy. This is evident from the diagonal values shown in Figures 4 and 6, where these percentages represent the comparison between expected and predicted objects for each class. These values are normalized, thereby testing accuracy on a scale from 0 to 1, where 0 indicates inaccurate detection and 1 signifies perfect detection. When the classes are compared separately, it can be seen that overall, the styled model has much lower values than the non-styled model. When we look at the classes, it is shown that the styled model scored 6% lower than the non-styled version. This percentage difference is even higher for some of the classes. For example, the belt has a difference of 12% when the two results are compared.

Furthermore, in order to conduct a detailed analysis of model accuracy and consistency, F1-Confidence curves are presented in Figures 5 and 7. These curves plot precision and recall values on the F1 score axis, comparing them across various confidence levels. The confidence level of 0.5 holds particular significance as it represents a threshold where the model's predictions are at an equilibrium of uncertainty—neither too confident nor too uncertain. Evaluating model performance at this point is crucial as it reveals how well the model discriminates between classes and handles ambiguous predictions. The comparison reveals that the styled model has significantly lower values in terms of accuracy at the 0.5 confidence level. When the comparison is made for all classes, the styled model has a lower F1 score. The difference in score is 0.15. This drop is seen in each of the classes. Additionally, it can be observed from the figures that the non-styled model is more consistent regardless

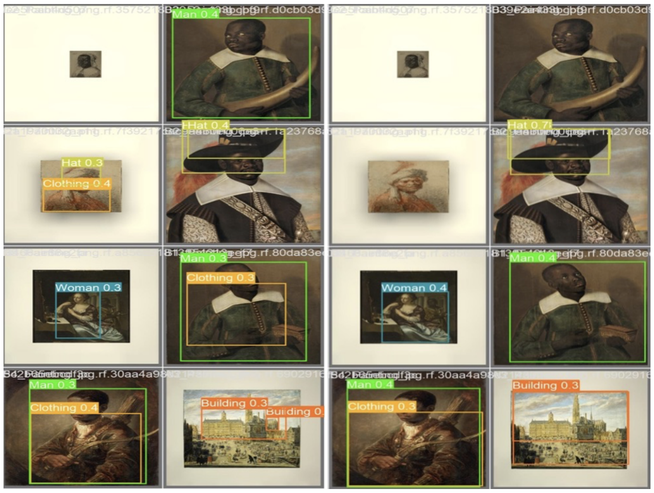


Fig. 8. Object detection results of the styled model on the left and non-styled model on the right for paintings.

of the chosen threshold. Thus, it is more accurate regardless of the chosen threshold.

The models are evaluated on the paintings to analyse the results, and as shown in Figures 8, the non-styled model has performed better on the paintings than the styled model. This observation is based on the number of instances that are detected in the non-styled model and not detected in the styled model. The styled model misses many more detection than the non-styled model. This observation was seen in all instances of the detection results, regardless of the threshold chosen for comparison. Additionally, if we look at the images and their corresponding boundary boxes, it is clear that the non-styled model performed more precisely than the styled model when it came to determining details. Especially the building detection shows the difference in precision. Both models have low confidence values for detecting objects, such as the confidence level shown for the class label "man." The confidence level is presented to be 0.3 for men. This observation shows that the collected datasets need to be more extensive, and the labelling of the dataset needs to be more precise to obtain a more accurate model. Additionally, this highlights the importance of further fine-tuning the models and improving data quality to achieve better performance in object detection tasks.

4.4 Eye Movement Patterns

Segmentation results are presented by overlaying the masks generated by the system onto the paintings. These results are produced using a pre-trained object detection model combined with a custom segmentation model. Each object is displayed in a different colour for visual distinction. An example of a segmented painting is shown in Figure 9. This image contains five objects, each coloured differently for better visualization. The objects identified are one man, one woman, one piece of clothing, one dress, and one table. These masks and labels are stored in a file to be further compared with the gaze information. Figure 10 represents an example result obtained from the comparison of the segmented painting shown in Figure 9.

The patterns are obtained from a single visitor session for all the paintings and the results are observed. These results of all the paintings for a single visitor showed that the visitor spent 37% of the session looking at the background rather than the meaningful objects. On the other hand, when the rest of the data is considered, it is shown that the visitor spent 30% of the time looking at "man", "woman" and their clothes. The second important observation was that the most interesting object was the "building" class for the visitor. This is observed from the fact that the visitor spent 23% of the session looking at the buildings. This shows that architecture is an important class to make a distinction while detecting objects. Lastly, the visitor spent 20% of the session looking at the remaining of the objects. Overall, the visitor looked at all the objects detected. However, considering the extensive time spent looking at the background, it can be observed that the 11 class labels defined in this research need to be increased through further analysis of the background spots observed by the visitors.

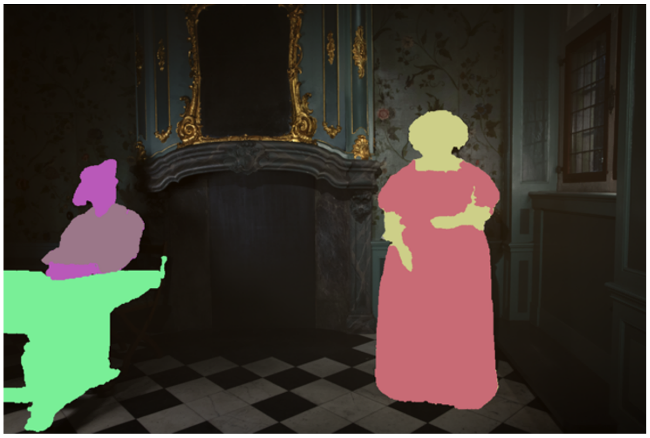
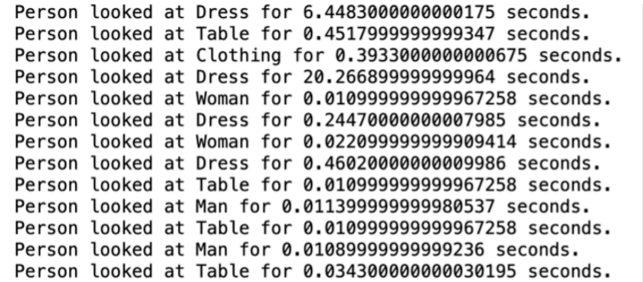


Fig. 9. An example plotted segmentation mask results on a painting.



```
Person looked at Dress for 6.4483000000000175 seconds.
Person looked at Table for 0.4517999999999347 seconds.
Person looked at Clothing for 0.3933000000000675 seconds.
Person looked at Dress for 20.266899999999964 seconds.
Person looked at Woman for 0.010999999999967258 seconds.
Person looked at Dress for 0.24470000000007985 seconds.
Person looked at Woman for 0.022099999999909414 seconds.
Person looked at Dress for 0.46020000000009986 seconds.
Person looked at Table for 0.010999999999967258 seconds.
Person looked at Man for 0.011399999999980537 seconds.
Person looked at Table for 0.010999999999967258 seconds.
Person looked at Man for 0.01089999999999236 seconds.
Person looked at Table for 0.03430000000030195 seconds.
```

Fig. 10. An example session representing the eye movement patterns of a visitor.

5 DISCUSSION

In this discussion, we explore the implications of our findings, provide the reasoning behind the limitations and unexpected results and supply the possible directions of future research.

The detailed findings from our study on the effects of the style-transfer method for object detection in paintings, combined with segmentation and eye gaze information, provided significant insights into identifying objects of interest in paintings. These insights into the eye movement patterns of visitors can be utilized further with other methods to enhance user engagement and increase attraction to VR museums.

The primary objective of this research was to determine the appropriate class labels that are significantly important for understanding the paintings. The analysis of the paintings and their descriptions showed that there are 11 important object categories to identify. These objects are mentioned in the descriptions to explain the paintings and their background like historical facts and influences. Later in the research, additional categories were identified as crucial because eye gaze data indicated that descriptions often lacked distinctions in categories such as dress and clothing. These categories were initially combined into a single category due to the lack of differentiation in descriptions. However, further analysis revealed their importance based on the eye gaze data. This finding suggests that descriptors should include more specific objects of interest in their explanations.

The secondary objective of this research was to determine whether style-transfer methods can enhance the accuracy of object detection models for classical paintings. This was assessed by analysing the validation results of two identical models trained on styled and non-styled datasets. The results indicated that the non-styled model had higher accuracy on a standard validation set compared to the styled model. This outcome was expected, as the purpose of styling was to improve accuracy on paintings, even if it meant losing accuracy on natural images. However, contrary to expectations, the styled model also performed better than the non-styled model when it came to detecting objects in paintings. This unexpected result could be attributed to several factors. First, the collected datasets were somewhat small for testing such scenarios, which might have influenced the outcomes. Another potential cause is that only one image was used for styling the dataset, leading to a more biased styled dataset rather than an unbiased one. Lastly, most of the classical paintings used in this research were portraits of people, which are more closely related to natural images rather than complex and highly stylized paintings. These are major factors to take into account. However, considering the time constraints and the lack of appropriate materials, the models performed sufficiently well to be utilized in the subsequent build of the system.

The last objective of this research was to combine the segmentation results with the eye gaze data to provide a piece of valuable information which can be used further in other studies to enhance user engagement and increase attraction to VR museums. This was achieved by developing a method that integrates object detection with segmentation to generate accurately labelled segmentation masks. Subsequently, these masks are compared with the eye gaze coordinates captured during the visitor's session in Unity, enabling the conversion of eye gaze coordinates into corresponding object labels. Furthermore, this label information is matched with timestamps to analyse visitor eye patterns. Eye movement patterns provide valuable information that can be utilized for numerous additional applications aimed at enhancing user engagement and visitor attraction. One use case of this information is to utilise these patterns for generating painting descriptions, which can be implemented with a virtual agent in the Unity environment to guide visitors. The guidance can provide information to visitors following the order obtained from the patterns. This method would allow visitors to easily follow the guidance and hopefully enhance their experience. Further user studies can be conducted using this method to analyse visitor patterns and establish a typical pattern per painting. This could provide a definitive framework for virtual agents to follow when describing each painting. Lastly, the pattern information obtained from this research provides an important result related to the chosen class labels for detection. The patterns are utilized to see which objects are more important to the visitors by analysing the timestamps. This analysis highlighted the importance of distinguishing between the man and the woman, as well as between dress and clothing. These distinctions significantly influence the painting's description and are crucial due to their connection with historical influences.

6 CONCLUSION

In summary, this research highlighted multiple avenues of integration of style transfer, object detection, segmentation and eye gaze analysis/understanding in unifying the experience, and interaction with paintings in VR museum environments. The results highlight the importance of the correct labelling and separation of object categories in paintings, as revealed by the visitors' eye gaze patterns. This view not only makes the explanation of paintings more vivid but also provides new ideas for the historical background.

In the study, we also examined how well style-transfer methods work for object detection in classical paintings. Nevertheless, the performance of the styled model was evaluated for object detection performance on paintings, and surprisingly, it showed competitive performance to foresee methods and to be useful in the cultural heritage domain, thus hinting at something that could be further explored and developed.

In addition, it presents an innovative method to combine segmentation masks and eye gaze data in supporting the study of visitor interactions with artworks, which provides benefits to the analysis of user engagement and the potential for improving the VR museum experience. Painting descriptions can be automatically generated regarding eye movement patterns, which underlines the practical relevance of this research for guided tours and individual visitor encounters. This research can be utilized to further advance the visitor engagement in VR art museums.

Overall, this study contributes to the evolving intersection of computer vision, human-computer interaction, and cultural heritage preservation, demonstrating how technological advancements can deepen our appreciation and understanding of art in immersive digital environments. It shows how these technologies can make art more accessible and engaging for everyone.

ACKNOWLEDGMENTS

During the preparation of this work, the author used ChatGPT in order to correct the grammar and the vocabulary of this paper. After using this tool/service, the author reviewed and edited the content as needed and took full responsibility for the content of the work.

REFERENCES

[1] C. Anthes, R. J. Garcia-Hernandez, M. Wiedemann, and D. Kranzlmuller. 2016. State of the art of virtual reality technology. In *2016 IEEE Aerospace Conference*. https://doi.org/10.1109/aero.2016.7500674

[2] C. Bailey-Ross, S. Gray, J. Ashby, M. Terras, A. Hudson-Smith, and C. Warwick. 2016. Engaging the Museum space: Mobilizing visitor engagement with digital content creation. *Digital Scholarship in the Humanities* 32, 4 (2016), 689–708. https://doi.org/10.1093/llc/fqw041

[3] F. Barth, H. Candello, P. Cavalin, and C. Pinhanez. 2020. Intentions, meanings, and whys. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. https://doi.org/10.1145/3405755.3406128

[4] N. Cohen, Y. Newman, and A. Shamir. 2022. Semantic segmentation in art paintings. *Computer Graphics Forum* 41, 2 (2022), 261–275. https://doi.org/10.1111/cgf.14473

[5] T. Diwan, G. Anirudh, and J. v. Tembhurne. 2023. Object detection using YOLO: challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications* 82, 6 (2023), 9243–9275. https://doi.org/10.1007/s11042-022-13644-y

[6] J. Dong, Q. Chen, S. Yan, and A. Yuille. 2014. Towards unified object detection and semantic segmentation. In *Computer Vision – ECCV 2014*. 299–314. https://doi.org/10.1007/978-3-319-10602-1_20

[7] B. Fasel and L. Van Gool. 2007. Interactive Museum guide: Accurate retrieval of object descriptions. In *Lecture Notes in Computer Science*. 179–191. https://doi.org/10.1007/978-3-540-71545-0_14

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2014.81

[9] M. A. Hashmani, M. M. Memon, and K. Raza. 2020. Semantic segmentation for visually adverse images – A critical review. In *2020 International Conference on Computational Intelligence (ICCI)*. https://doi.org/10.1109/icci51257.2020.9247758

[10] K. He, G. Gkioxari, P. Dollar, and R. Girshick. 2017. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*. 2961–2969. https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html

[11] D. Javdani Rikhtehgar, S. Wang, H. Huitema, J. Alvares, S. Schlobach, C. Rieffe, and D. Heylen. 2023. Personalizing cultural heritage access in a virtual reality exhibition: A user study on viewing behavior and content preferences. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. https://doi.org/10.1145/3563359.3596666

[12] M. Kejriwal. 2019. What is a knowledge graph? Domain-Specific Knowledge Graph Construction. In *Domain-Specific Knowledge Graph Construction*. 1–7. https://doi.org/10.1007/978-3-030-12375-8_1

[13] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar. 2019. Panoptic Segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 9404–9413. https://openaccess.thecvf.com/content_CVPR_2019/html/Kirillov_Panoptic_Segmentation_CVPR_2019_paper.html

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollar, and R. Girshick. 2023. Segment Anything. In *International Conference on Computer Vision (ICCV)*. 4015–4026. https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html

[15] H. Lee, T. H. Jung, M. Tom Dieck, and N. Chung. 2020. Experiencing immersive virtual reality in museums. *Information & Management* 57, 5 (2020), 103229. https://doi.org/10.1016/j.im.2019.103229

[16] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen. 2020. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision* 128, 2 (2020), 261–318. https://doi.org/10.1007/s11263-019-01247-4

[17] S. Machała, N. Chamier-Gliszczyński, and T. Królikowski. 2022. Application of AR/VR technology in industry 4.0. *Procedia Computer Science* 207 (2022), 2990–2998. https://doi.org/10.1016/j.procs.2022.09.357

[18] M. M. Memon, M. A. Hashmani, A. Z. Junejo, S. S. Rizvi, and K. Raza. 2022. Unified DeepLabV3+ for semi-dark image semantic segmentation. *Sensors* 22, 14 (2022), 5312. https://doi.org/10.3390/s22145312

[19] P. P. Morantes, S. A. Penarete, G. Arbelaez, M. Camargo, and L. Dupont. 2016. Understanding Museum visitors' experience through an eye-tracking study and a living lab approach. In *2016 International Conference on Engineering, Technology and Innovation/IEEE lnternational Technology Management Conference (ICE/ITMC)*. https://doi.org/10.1109/ice/itmc39735.2016.9025900

[20] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2014.119

[21] L. P. Osco, Q. Wu, E. L. de Lemos, W. N. Gonçalves, A. P. Ramos, J. Li, and J. Marcato. 2023. The segment anything model (SAM) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation* 124 (2023), 103540. https://doi.org/10.1016/j.jag.2023.103540

[22] R. Padilla, S. L. Netto, and E. A. B. da Silva. 2020. A Survey on Performance Metrics for Object-Detection Algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130

[23] J. B. Schreiber, A. J. Pekarik, N. Hanemann, Z. Doering, and A. Lee. 2013. Understanding visitor engagement and behaviors. *The Journal of Educational Research* 106, 6 (2013), 462–468. https://doi.org/10.1080/00220671.2013.833011

[24] L. Sun, K. Wang, K. Yang, and K. Xiang. 2019. See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*. https://doi.org/10.1117/12.2532477

[25] F. Umam, F. Adiputra, A. Dafid, and S. Wahyuni. 2022. Autonomous Museum tour guide robot with object detection using TensorFlow learning machine. In *2022 IEEE 8th Information Technology International Seminar (ITIS)*. https://doi.org/10.1109/itis57155.2022.10009997

[26] D. Wang and D. He. 2022. Fusion of mask RCNN and attention mechanism for instance segmentation of apples under complex background. *Computers and Electronics in Agriculture* 196 (2022). https://doi.org/10.1016/j.compag.2022.106864

[27] P. Weinzaepfel, G. Csurka, Y. Cabon, and M. Humenberger. 2019. Visual localization by learning objects-of-Interest dense match regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2019.00578

[28] J. Whyte. 2003. Innovation and users: Virtual reality in the construction sector. *Construction Management and Economics* 21, 6 (2003), 565–572. https://doi.org/10.1080/0144619032000113690

[29] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun. 2019. UPSNet: A unified panoptic segmentation network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2019.00902

[30] X. Ying, B. Lang, Z. Zheng, and M. C. Chuah. 2022. Delving into light-dark semantic segmentation for indoor scenes understanding. In *Proceedings of the 1st Workshop on Photorealistic Image and Environment Synthesis for Multimedia Experiments*. https://doi.org/10.1145/3552482.3556556

[31] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun, and Y. Tang. 2018. Methods and datasets on semantic segmentation: A review. *Neurocomputing* 304 (2018), 82–103. https://doi.org/10.1016/j.neucom.2018.03.037

[32] X. Yue, K. Qi, X. Na, Y. Zhang, Y. Liu, and C. Liu. 2023. Improved yolov8-seg network for instance segmentation of healthy and diseased tomato plants in the growth stage. *Agriculture* 13, 8 (2023). https://doi.org/10.3390/agriculture13081643

[33] S. Zhang, Z. Zhang, L. Sun, and W. Qin. 2019. One for all: A mutual enhancement method for object detection and semantic segmentation. *Applied Sciences* 10, 1 (2019), 13. https://doi.org/10.3390/app10010013